IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 26, 2021, accepted November 16, 2021, date of publication November 19, 2021,
date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129480

CRF-EfficientUNet: An Improved UNet Framework for Polyp Segmentation in Colonoscopy Images With Combined Asymmetric Loss Function and CRF-RNN Layer

LE THI THU HONG[1], NGUYEN CHI THANH[1], AND TRAN QUOC LONG[2]
[1]Institute of Information Technology, AMST, Hanoi 840000, Vietnam
[2]Faculty of Information Technology, University of Engineering and Technology, VNU, Hanoi 700000, Vietnam

Corresponding author: Nguyen Chi Thanh (thanhnc80@gmail.com)

ABSTRACT Colonoscopy is considered the gold-standard investigation for colorectal cancer screening. However, the polyps miss rate in clinical practice is relatively high due to different factors. This presents an opportunity to use AI models to automatically detect and segment polyps, supporting clinicians to reduce the number of polyps missed. Inspired by the success of UNets, a popular strategy for solving medical image segmentation tasks, this article proposes a novel framework for polyp segmentation called CRF-EfficientUNet, which enhances UNet using the EfficientNet encoder, a combined asymmetric loss function, and Conditional Random Field as a Recurrent Neural Network (CRF-RNN) layer on top. A novel loss function that combines pixel-wise cross-entropy loss and asymmetric similarity loss to solve the unbalanced imaging data problem is proposed. Training the proposed network with this loss function can achieve a considerably higher Dice score and better polyp segmentation prediction. In addition, we add the CRF-RNN layer to the proposed framework to improve the quality of semantic segmentation. Experimental results on popular benchmark datasets show that CRF-EfficientUNet achieves state-of-the-art accuracy compared to existing methods. The results of the experiments, which are performed on the CVC-ClinicDB dataset for training and testing, are 95.55% Dice and 92.23% IoU. While the experimental results on cross-dataset using Kvasir-SEG as the training set, CVC-ColonDB as the test set are 85.59% Dice and 76.19% IoU. These results indicate that the proposed method has high generalization capability and learning ability, and it can be a compelling choice for practical applications with considerable data variations. The source code is available at: https://github.com/lethithuhong1302/CRF-EfficientUNet

INDEX TERMS Polyp segmentation, medical image analysis, deep learning, loss function.

I. INTRODUCTION

Colorectal cancer (CRC) is one of the most common causes of cancer-related death in the world for both men and women, with 576,858 deaths (account for 5.8% of all cancer deaths) worldwide in 2020 [1]. CRC usually arises from abnormal polyp growth inside the colon, although polyps grow slowly and may take years to become cancer. According to anatomical findings, the structure of polyps is distinguished from normal mucosa by color, size, and surface type. The surface of polyps can be flat, elevated, or pedunculated based on a change in the gastrointestinal tract [2]. Though not all polyps lead to CRC, all CRC starts with polyps that become cancerous over time. While the advanced stages of colorectal cancer have a poor five-year survival rate of 10%, the early diagnosis has shown a significantly more favorable five-year survival rate of 90% [3]. Therefore, accurate detection, investigation, and analysis of types, patterns, and structures of polyps are important to reduce the spread of CRC. Nowadays, colonoscopy is considered the primary method for colon screening and preventing polyps from becoming cancerous. However, colonoscopy suffers from human errors because it depends on highly skilled endoscopists and a high level of eye-hand coordination. Moreover, some of the rare types of polyps are visually difficult to distinguish due to flat natures that demand the experiences and expertise of endoscopists.

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

Previous studies confirmed that 22%–28% of polyps are missed in patients undergoing colonoscopy [4]. Segmenting out polyps from the normal mucosa can help endoscopists to improve their segmentation errors and subjectivity. Therefore, this study focuses on the polyp segmentation problem using deep learning methods.

This study focuses on the polyp segmentation problem using deep learning methods. Precise segmentation of the polyp regions is particularly complicated because polyps have different shapes, sizes, colors, and appearances [5]. In addition, there are challenges such as the presence of image artifacts like specularity, saturation, artifact, bubbles, and instrument [6], intestinal contents, and low-quality images that can cause errors during segmentation. Figure 1 shows some of the challenges presented by colonoscopy images. Over the past years, researchers have made several efforts to develop Computer-Aided Diagnosis(CADx) prototypes for automated polyp segmentation. Most of the prior polyp segmentation approaches were based on analyzing polyp color, texture, shape, or edge information to segment polyp regions. More recently, deep neural networks have been widely used to solve medical image segmentation problems, including polyp segmentation. The CADx system for automatically segmenting out polyps from normal mucosa on colonoscopy images can be an effective clinical tool that helps endoscopists for faster screening and higher accuracy [5]. For building a powerful polyp segmentation CADx system that could be used in clinical settings, it is necessary to address two common challenges: (i) Robustness (i.e., the ability of the system to perform well on both easy and challenging images), and (ii) Generalization (i.e., a system trained on a dataset from a specific hospital should generalize across different hospitals) [7]. To address the aforementioned research challenges and issues, the overall goal of this article is to develop a novel deep learning framework for polyp segmentation with high generalizability and learning ability, so that it can be an effective choice for practical applications.

Among various deep learning models, UNet [8] and its variants have demonstrated impressive performance in biomedical image segmentation. Motivated by the success of UNet, in this work, we propose a novel polyp segmentation method based on the UNet architecture. We adapt the UNet model for polyp segmentation and aim to evaluate the model with different encoders (MobileNet [9], ResNet [10], and EfficientNets [11]). We choose the EfficientNetB7 encoder for our model because of the highest performance. One of the challenges in training networks for polyp segmentation is unbalanced data, i.e., polyp pixels are often much lower in numbers than non-polyp pixels. Networks trained by unbalanced data may make predictions with high precision and low recall. These predictions are severely biased toward the non-polyp class, which are particularly undesired because the consequences of false negatives would be more serious than those of false positives. Therefore, we propose a novel loss function that combines pixel-wise cross-entropy loss and asymmetric similarity loss for training

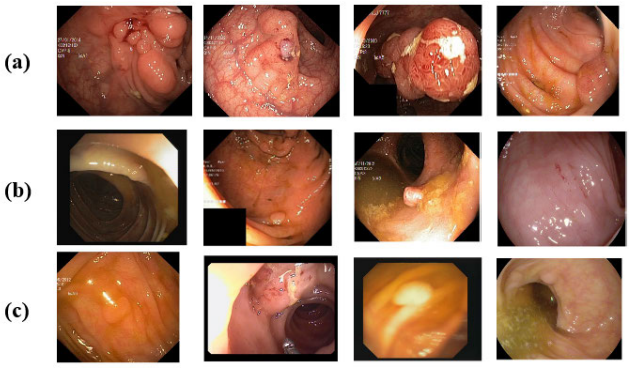

FIGURE 1. Some of the challenges presented by colonoscopy images: (a) Varying shapes and textures of polyps, (b) Small polyps, (c) Blurriness, intestinal contents, flares, or low-quality images.

polyp segmentation models to address this problem. By training models with the proposed loss function, we found that the models can make predictions with a better trade-off between precision and recall prediction to yield accurate polyp segmentation. Moreover, one central issue in polyp segmentation is the limited capacity of deep learning techniques to delineate polyp objects. To solve this problem, we use a deep network that fully integrates Conditional Random Fields (CRFs) [12] probabilistic graphical modeling with CNN, making it possible to train the whole deep network end-to-end with the back-propagation algorithm, avoiding offline post-processing methods for object delineation [13]. Finally, we perform experiments on a range of recent public datasets for polyp segmentation, i.e., Kvasir-SEG [14], CVC-ClinicDB [15], CVC-ColonDB [16], EITS-Larib [15] with different scenarios of using training and test data to evaluate our proposed method and compare with state-of-the-art (SOTA) approaches.

This article is an extension of our work originally presented in the 2020 RIVF International Conference on Computing and Communication Technologies (RIVF) [17]. We extend previous work by (i) modified the model architecture by remove the ensemble step and add a CRF-RNN layer, (ii) use EfficientNetB7 instead of EfficientNetB5 as encoder, (iii) conducted comprehensive experiments with multiple datasets, multiple experiment settings for comparison with recent SOTAs in polyp segmentation and ablation study. In summary, this article makes the following key contributions:

1) We present a novel neural network architecture for automatic polyp segmentation, called CRF-EfficientUNet, extended from UNet architecture with an EfficientNet encoder and CRF-RNN layer on top. Moreover, we use the transfer learning method on the proposed network architecture to achieve better performance.

2) We propose a loss function that combines pixel-wise cross-entropy loss and asymmetric similarity loss called the combined asymmetric loss function for training polyp segmentation networks. The combined asymmetric loss function can effectively boost the performance of polyp segmentation

networks. The proposed loss function was used to train our polyp segmentation model results in a better performance.

3) We train and validate the proposed method on four popular benchmark datasets, i.e., Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, EITS-Larib, with different scenarios of using training and testing data. The results show that our model has the robustness to detect small polyps that are frequently missed during colonoscopy and perform well on easy images. Moreover, our network CRF-EfficientUNet outperforms all SOTAs across unseen polyp datasets; this demonstrates that our proposed method has better generalizability than existing methods. The experimental results indicate that the proposed model can be a compelling choice for practical applications with considerable data variations.

The rest of the paper is organized as follows. Section II reviews related research on polyp segmentation. Section III describes the proposed method for polyp segmentation in detail. Section IV outlines our experiment settings. The experimental results and discussion are presented and discussed in Section V. Finally, Section VI summarizes and concludes this work.

II. RELATED WORK

Many methods have been proposed that focus on accurate polyp segmentation. The existing research works in polyp segmentation can be roughly grouped into main approaches: using image processing segmentation and traditional machine learning methods, and using deep learning methods. The processing segmentation methods analyze either the polyp's edge or its color and texture for polyp segmentation. Bernal *et al.* [16] proposed to use the ''depth of valleys'' of an image to segment polyps. They use the watershed algorithm to segment images into polyp candidate regions and then classify each region into polyp and non-polyp. This classification is based on region information and the ''depth of valleys'' in each region. Ganz *et al.* [18] propose a method based on Hough transform to detect the region of interest (ROI) and specular reflection suppression with an exemplar-based image in painting as a preprocessing method. Then, they use an algorithm called shape-UCM for image segmentation, shape-UCM works based on image gradient contours and spectral clustering. Traditional machine learning methods are based on hand-crafted features for image representation. These methods use color, texture, shape, or edge information as extracted features and train the classifier to distinguish polyps from surrounding normal mucosa. Tajbakhsh *et al.* [19] proposed a feature extraction method to extract sub-patches with a 50% overlap and calculate their average vertically, resulting in one-dimensional signals. After that, they use DCT coefficients as a feature for each extracted patch. Finally, they use a two-stage random forest classifier to label each patch.

The deep learning-based approach for polyp segmentation has gained much attention in recent years due to the automatic feature extraction process to segment polyp regions with unprecedented precision. In addition, the public database of polyp images facilitated further research on the use of deep learning models for polyp segmentation. Qadir *et al.* [20] proposed using Mask-RCNN incorporated with traditional CNN-based feature extractors to provide bounding boxes of the polyp regions. Kang and Gwak [21] used Mask-RCNN, which relies on ResNet50 and ResNet101, as a backbone structure for automatic polyp detection and segmentation. For obtaining pixel-level segmentation, a fully convolutional neural network (FCN) was used. The authors in [22] showed that FCN architectures could be refined and adapted to recognize polyp structures. Zhang *et al.* [23] used FCN-8S to segment polyp region candidates, and texton features computed from each region were used by a random forest classifier for the final decision. Fan *et al.* [24] propose PraNet, enhancing an FCN-like model using a parallel partial decoder and reverse attention modules for medical image segmentation. Instead of a single encoder in traditional FCN architecture, UNet is proposed, which increases the performance of FCN considerably and has established itself as a popular choice in medical image segmentation. UNet is an encoder-decoder-based structure that uses skip connections to concatenate the features from the encoding and decoding layers. Inspired by the success of UNet, several variants were proposed for polyp segmentation and yielded promising results. Jha *et al.* [25] present DoubleU-Net, which combines two UNets. The first UNet uses a pre-trained VGG-19 as the backbone. The second UNet is added at the bottom of the first UNet to capture more semantic information efficiently. They also adopt Atrous Spatial Pyramid Pooling (ASPP) to capture contextual information within the network. Zhou *et al.* [26] propose UNet++, a deeply supervised encoder-decoder network, which connects UNets through a series of nested, dense skip pathways. Jha *et al.* [27] also propose ResUNet++, which takes advantage of residual blocks, squeeze and excitation units, ASPP, and the attention mechanism. Similar to UNet, another deep convolutional encoder-decoder architecture, Segnet [28], is also used for polyp segmentation. Wang *et al.* [29] used the SegNet architecture to detect polyps in real-time and with high sensitivity and specificity. Afify *et al.* [30] presented an improved framework for polyp segmentation based on image preprocessing and two types of SegNet architecture. Mahmud *et al.* [31] proposed PolypSegNet, a modified SegNet architecture for automated polyp segmentation from colonoscopy images with several sequential depth dilated inception (DDI) blocks, deep fusion skip modules (DFSM), and deep reconstruction module (DRM). Additionally, there are several recent studies on polyp segmentation [32]–[35]. They are useful steps toward building an automated polyp segmentation system.

From the presented related works, we observe that works on polyp segmentation problems are becoming mature. Researchers are conducting a variety of studies with many different methods for precision polyp segmentation. However, the main drawback in the field is that very few works apply towards testing the generalizability of models with the cross-dataset test. Most of the current works have proposed

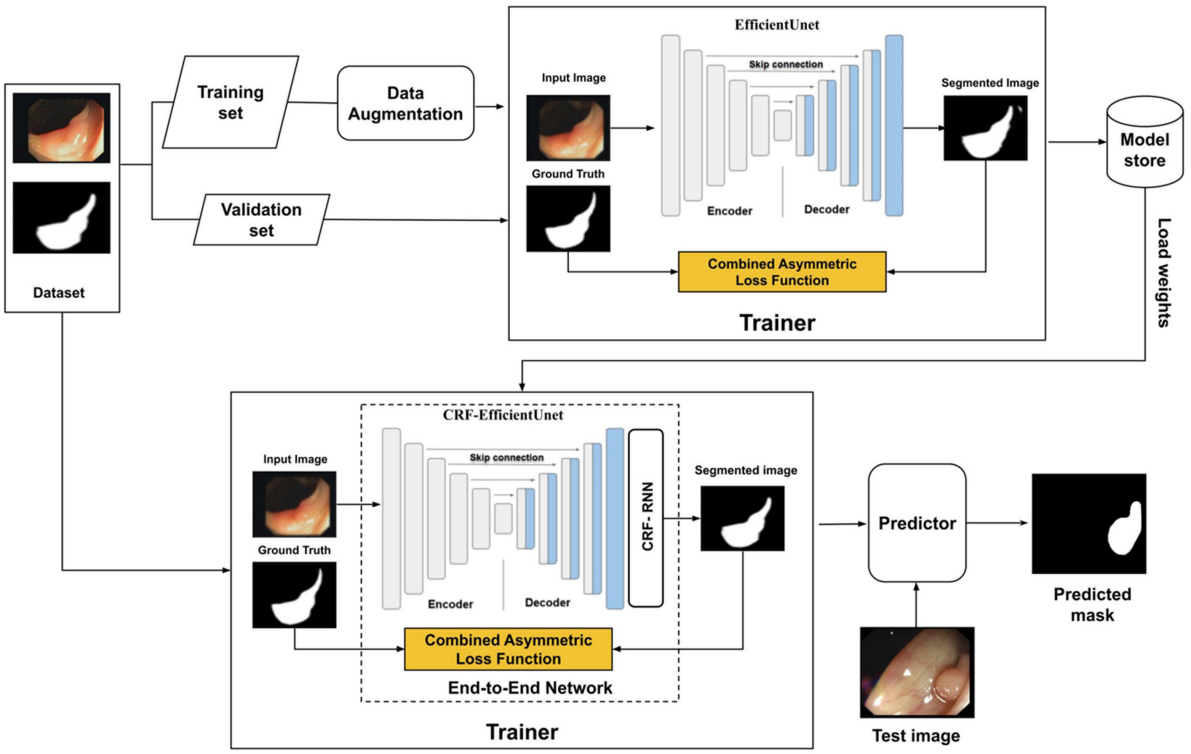

FIGURE 2. Flowchart of the proposed deep learning model for polyp segmentation.

algorithms tested on single, often small, imbalanced, and explicitly handpicked datasets. Besides, many challenging polyps are usually missed during colonoscopy examinations and can develop into cancer if they are not detected early. Moreover, one of the significant challenges in the medical domain is the lack of large training datasets, and the obtained datasets are often imbalanced. These challenges make it harder to build robust and generalizable systems for precision polyp segmentation. Toward addressing these challenges, in this work, we aim to develop an algorithm that could achieve high performance on different datasets. We have done extensive experiments on various colonoscopy images. Furthermore, we have trained the proposed model on datasets from multiple clinical settings and tested it on other diverse unseen datasets to achieve the goal of building generalizable and robust models.

III. METHODOLOGY

A. OVERVIEW OF THE PROPOSED METHOD

The overall architecture of our proposed network, CRF-EfficientUNet, is depicted in Figure 2. First, we evaluate the performance of the UNet architecture for polyp segmentation with different CNN encoders. We select the EfficientNet B7 encoder for the UNet architecture due to it gives the highest performance. Next, we extended the UNet architecture with the EfficientNet B7 encoder and a CRF-RNN layer on top. Besides, we propose a novel loss function that combines pixel-wise cross-entropy loss and asymmetric similarity loss called the combined asymmetric loss function. Training the networks uses combined asymmetric loss, and the transfer learning method can effectively boost the network's segmentation performance. The CRF-RNN layer is integrated on top of UNet as follows. First, EfficientUnet was trained. When the UNet network's parameters have been trained, they are fixed and set to untrainable. Next, the softmax layer is left out, and the CRF-RNN layer on top is integrated. Finally, the CRF-EfficientUNet is trained end-to-end once again.

B. UNets WITH DIFFERENT ENCODERS FOR POLYP SEGMENTATION

The UNet architecture was developed by Ronneberger *et al.* for Biomedical Image Segmentation [8]. UNet has two symmetric paths. The first path is also called the encoder, which is used to capture the context in the image. The encoder consists of convolutional and max-pooling layers. The second is called the decoder, which is used to enable precise localization using transposed convolutions. Moreover, UNet has connections between encoder and decoder to skip the higher-level features the encoder learned that could be lost during the decoding process. That means the outputs of the encoding layers are passed directly to the decoding layers so that all the important pieces of information can be preserved. We adopt a transfer learning approach with UNet architecture for polyp segmentation. We use UNet with a CNN model

pre-trained on the ImageNet dataset as the encoder. The choice of the encoder is essential because the CNN architecture, the number of parameters, the type of layers directly affect the speed, memory usage, and most importantly, the performance of the UNet. In this work, we select three architectures to compare and evaluate their performance in polyp segmentation: MobileNet [9], ResNet [10], and EfficientNet [11]. MobileNet is a family of mobile-first computer vision models from Google. They are designed to maximize accuracy while being mindful of the restricted resources for an on-device or embedded application. ResNet is a residual learning framework that enables training deep networks easily. With ResNet, we can benefit from deeper CNN networks to obtain an even higher level of essential features for challenging tasks such as polyp segmentation. EfficientNets are the latest family of image classification models from Google, which achieves the state of the art accuracy on ImageNet. Mingxing Tan and Quoc V. Le proposed the EfficientNets based on AutoML and Compound Scaling. In particular, they use the AutoML MNAS Mobile framework to develop a mobile-size baseline network named EfficientNet-B0. Then, they use the compound scaling method to scale up this baseline to obtain EfficientNet-B1 to EfficientNet-B7. The accuracies of networks are steadily increasing while maintaining a relatively small size from EfficientNet-B0 to EfficientNet-B7. This study conducts an ablation study on different encoders, including EfficientNets family from EfficientNet-B0 to EfficientNet-B7, ResNet-50, ResNet-101, and MobileNetV2. Our experiments show that UNet with EfficientNet B7 encoder gives the highest performance.

C. COMBINED ASYMMETRIC LOSS FUNCTION

We present the combined asymmetric loss function, a novel loss function that combines existing loss functions with hyper-parameters to boost segmentation results: cross-entropy loss and asymmetric similarity loss. Pixel-wise cross-entropy loss was used by Ronneberger *et al.* in [8] for the task of image segmentation. This loss simply verified each pixel individually, comparing the class predictions defined as a depth-wise pixel vector to the target vector. The cross-entropy loss function is defined as:

$$\mathcal{L}_{CE} = \sum_{i,j} g_{i,j} * \log\left(p_{i,j}\right) \tag{1}$$

where $p_{i,j}$ is the predicted segmentation probability, and $g_{i,j}$ stands for the ground truth at image pixel (i, j). The cross-entropy loss function assesses every single pixel. In medical imaging applications, such as polyp segmentation, the polyp pixel class is much lower in number than the none-polyp pixel class. Hence, the segmentation network trained with a cross-entropy loss function is biased towards the background image than the object itself. Furthermore, as the foreground region is often missing or only partially detected, it is not easy for the model to see the object.

Dice score coefficient *(DSC)* is an overlap index widely used to assess segmentation maps in the medical community.

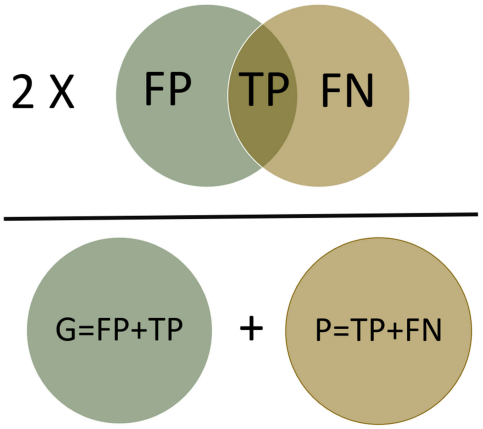

FIGURE 3. Dice score: TP is true positives, FP is false positives, and FN is false negatives, P is the set of predicted binary labels, G is the set of ground truth binary labels.

Dice similarity coefficient between the set of predicted binary labels (denoted as P) and the set of ground truth binary labels (denoted as G) is defined as:

$$DSC(P, G) = \frac{2|PG|}{|P| + |G|} \tag{2}$$

The dice loss function is formulated based on the Dice score [36]. This is used to improve UNet and other segmentation networks training. Simply put, Dice score is *2 × the area of overlap between P (predicted area) and G (ground truth area) divided by the total number of pixels in P and G*. Figure 3 illustrates the Dice score. We can see that all the Dice score considered is the foreground class but no background class. In other words, no matter how many ground-truth background pixels exist in the image, they will not affect the calculation of the Dice score. So that, The Dice score drops sharply when much more ground-truth background pixels than the ground-truth foreground pixels. That means the Dice score gets more sensitive when the image suffers severe class unbalanced. Moreover, as Figure 3, the Dice score can be calculated as:

$$DSC = \frac{2|TP|}{2|TP| + |FP| + |FN|} \tag{3}$$

where TP is true positives, FP is false positives, and FN is false negatives. In this equation, Dice score weighs false positives (FPs) and false negatives (FNs) equally. When data is class-imbalanced, positive (polyp) pixels are often much lower in numbers than negative (non-polyp) pixels. The network trained with Dice loss on imbalanced data may make predictions severely biased towards the negatives (non-polyp) class. That is particularly undesired in colonoscopy scan applications where false negatives are more serious than false positives. On the other hand, precision and recall are defined as:

$$Re = \frac{|TP|}{|TP| + |FN|} \tag{4}$$

$$Pre = \frac{|TP|}{|TP| + |FP|} \tag{5}$$

Combine Equation (3),(4),(5), we have:

$$DSC = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

As Equation 6, Dice score is the harmonic mean of precision and recall. A trained network with Dice loss on unbalanced data may make predictions with high precision and low recall. In some fields like medical image segmentation problems, however, the data are highly unbalanced, detecting the small number of pixels in the positive class is important. Thus, it is necessary to better balance precision and recall in training segmentation networks for unbalanced data. Asymmetric similarity loss function was proposed in [37] for training segmentation networks to make a better balance between precision and recall. The asymmetric similarity loss function is based on F_β score and used to replace Dice loss function. F_β score is defined as:

$$F_\beta = (1 + \beta^2)\frac{precision * recall}{\beta^2 * precision + recall} \tag{7}$$

By changing the hyperparameter β, we can control the trade-off between precision and recall. Equation 7 can be written as:

$$F(P, G, \beta) = \frac{(1 + \beta^2)|PG|}{(1 + \beta^2)|PG| + \beta^2|G\backslash P| + |P\backslash G|} \tag{8}$$

where $|P\backslash G|$ is the difference of P and G. Therefore, F_β score can be calculated as follows:

$$F_\beta = \frac{(1 + \beta^2)\sum p_{i,j}g_{i,j}}{(1+\beta^2)\sum p_{i,j}g_{i,j}+\beta^2\sum(1-p_{i,j})g_{i,j}+\sum p_{i,j}(1-g_{i,j})} \tag{9}$$

F_β score with the hyper-parameter β generalizes Dice similarity coefficient and Jaccard (IoU) index. When $\beta = 1$, the F_β score is Dice score, $\beta = 2$ generates F2 score, and $\beta = 0$ transforms the score to precision. When the hyper-parameter β is larger, the weight of recall is higher than the weight of precision, and the false negatives are more emphasized.

In this work, we proposed a combined asymmetric loss function that combines cross-entropy loss and asymmetric similarity loss for training networks to boost polyp segmentation results. The proposed loss function is defined as:

$$\mathcal{L}_{AsymCE} = \alpha * \mathcal{L}_{CE} + \mathcal{L}_{Asym} \tag{10}$$

where $\mathcal{L}_{CE}$ is cross-entropy loss and $\mathcal{L}_{Asym} = 1 - F_\beta$ is asymmetric similarity loss which is based on F_β score, the hyperparameter α controls the amount of cross-entropy loss term contribution in the loss function. Due to the polyp segmentation problem is also a pixel classification problem, we use the cross-entropy loss term to verified each pixel individually. However, cross-entropy loss assesses every single pixel. In colonoscopy images, polyps usually have a small surface area. Hence, the segmentation network trained with a cross-entropy loss function is biased towards the background rather than the polyp objects. Like Dice loss, asymmetric similarity loss can handle the input class-imbalance problem, e.g., segmenting small polyps from a large background. Moreover, asymmetric similarity loss allows training networks that make a better balance between precision and recall. By combining cross-entropy loss and asymmetric similarity loss for training networks, we can leverage the asymmetric similarity loss term to handles the input class-imbalance problem and control the trade-off between precision and recall. At the same time, we can force networks to learn better parameters by penalizing for false positives/negatives using the cross-entropy loss term. In the proposed loss function, appropriate values of the hyper-parameter α, β can be defined based on class imbalance ratios of the dataset. Our experimental results prove that combined asymmetric loss function is more robust than cross-entropy loss function and Dice loss function.

Table 1 lists recent polyp segmentation work that used different loss functions for training models. As reported in the table, none of the current loss functions can explicitly handle all the main challenges in the polyp segmentation problem. These challenges are handling class imbalance, the trade-off between precision and recall, and penalizing for false positives and false negatives. Some studies attempted to deal with class imbalance by using variants of cross-entropy loss and Dice loss: Sánchez-Peralta *et al.* [39] use a loss function that combines binary cross-entropy and Jaccard index loss; Nguyen *et al.* [40] use an adaptive weighted loss function which is a weighted cross-entropy loss; Mahmud *et al.* use Modified Focal Tversky (MFTL) loss function for training the PolypSegNet, MFTL increase the focus on hard training samples by utilizing Tversky index (a generalization of Dice score). However, these methods on polyp segmentation datasets do not handle the trade-off between precision and recall. Nguyen and Lee in [38] also proposed a loss function that combines the binary cross-entropy and the Dice loss. Milletari *et al.* in [37] also proposed a loss function that combines the binary cross-entropy and the Dice loss. Their loss function could penalize false positives and false negatives. But the trade-off between precision and recall couldn't be dealt with on all polyp segmentation test sets. In this article, we propose a combined asymmetric loss function that combines cross-entropy loss and asymmetric loss to train our polyp segmentation model. When the proposed loss function is used as an optimization function, the polyp segmentation model can handle class imbalance, the trade-off between precision and recall, and penalize for false positives and false negatives. The experiment's results in section V-A2 show that when our model CRF-EfficientUnet was trained with combined asymmetric loss function, and it significantly improves the polyp segmentation accuracy.

D. INTEGRATING CRF AS RNN LAYER ON TOP OF THE POLYP SEGMENTATION NETWORKS

Using a fully connected Conditional Random Field (CRF) in conjunction with a deep segmentation model is the popular approach for semantic segmentation. The idea behind this is that the segmentation model plays a role as a feature extractor

TABLE 1. Applied loss functions in the existing deep models for polyp segmentation.

Method	Loss function	Trade-off (pre and rec)	Handling class (imbalance)	Penalizing (for FPs/FNs)
ResUNet++ [28]	Dice loss	No	No	No
DoubleU-Net [26]	Binary cross-entropy loss	No	No	Yes
CDED-net [39]	Combination of binary cross-entropy and Dice loss	Yes	No	Yes
Sánchez-Peralta et al [40]	Combination of binary cross-entropy and Jaccard index	-	Yes	Yes
MED-Net [41]	Adaptive weighted loss function	No	Yes	Yes
A-DenseUNet [35]	Binary cross-entropy loss	No	No	Yes
PolypSegNet [32]	Modified Focal Tversky loss	No	Yes	No
Debesh Jha et al [34]	Binary cross-entropy loss	No	No	Yes
Efficient U-Net multi-scale attention [42]	Binary cross-entropy loss	No	No	Yes
Proposed method	**Combination of cross-entropy and asymmetric loss function**	**Yes**	**Yes**	**Yes**

that produces a coarse segmentation. Then CRF refines the result segmentation. The input of CRF includes the segmentation probality produced by the network and the original input image. Unlike a convolution layer that implements local filters, the fully-connected CRF considers every possible pair of pixels in the image. Each pair is called a clique. In CRF graphical model, the clique is defined by the spatial distance and color distance between pixels. This makes segmentations produced by the CRF much sharper than those produced by the original segmentation model. Thus, the receptive field of a CRF is the entire image. However, when using a CRF to improve the quality of a segmentation model, the CRF has to be trained separately after the base model has been trained. Hence, in [13], the authors propose the CRF mean-field approximation as Recurrent Neural Network (RNN) that can be added on top of CNN and train the whole system end-to-end.

In the fully connected pairwise CRF model, the image segmentation problem is solved as an optimization problem by minimizing an energy function [12]:

$$E(Y) = \sum_{i=1}^{N} \Phi(y_i^u) + \sum_{\forall i,j, i<j} \Psi(y_i^u, y_j^v) \tag{11}$$

The term $\Phi(y_i^u)$ measures the cost of assigning label u to pixel i, N is the number of pixels in the image, the pairwise potential $\Psi(y_i^u, y_j^v)$ measures the cost of assigning label u and v jointly to pixel i, j and is defined as:

$$\Psi(y_i^u, y_j^v) = \mu(u, v) \sum_{m=1}^{K} w^{(m)} k^{(m)}(f_i, f_j) \tag{12}$$

where $\mu(u, v)$ indicates the compatibility of labels u and v, $K = 2$ is the number of Gaussian kernels; $k^{(m)}$ is a Gaussian kernel, $w^{(m)}$ is a weight for the Gaussian kernel,f_i, f_j denote feature vectors of pixels i, j respectively.

$$k^{(1)} = \left(-\frac{|s_i - s_j|}{2\theta_\alpha^2} - \frac{|e_i - e_j|}{2\theta_\beta^2} \right) \tag{13}$$

$$k^{(1)} = \exp\left(-\frac{|s_i - s_j|}{2\theta_\gamma^2} \right) \tag{14}$$

where e_i, e_j denote the intensity and s_i, s_j denote spatial coordinates of pixels i, j respectively; $\theta_\alpha, \theta_\beta, \theta_\gamma$ are parameters

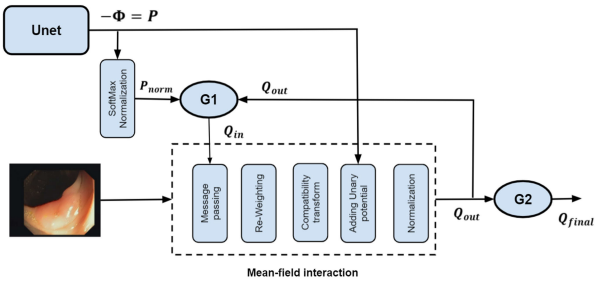

FIGURE 4. Network structure of CRF-RNN.

of the Gaussian kernels. Fully connected CRF predicts the probability of assigning label u to pixel i (q_i^u) by minimizing Equation (11). $\{q_i^u\}$ can be calculated using a mean-field iteration algorithm which is formulated as Recurrent Neural Networks. So that, CNNs and the fully connected CRF are integrated as one deep network and can be trained using a back-propagation [13].

This article presents a deep learning model that integrates UNet and CRF-RNN for polyp segmentation that can be trained the whole system end-to-end. Figure 4 shows the network structure of CRF-RNN in our proposed model. In Figure 4, G1, G2 are two gating functions:

$$Q_{in} = \begin{cases} softmax(P) & \text{if } t = 0 \\ Q_{out} = MeanField(Q_{in}) & \text{if } 0 < t \leqslant T) \end{cases} \tag{15}$$

$$Q_{final} = \begin{cases} 0 & \text{if } 0 < t < T \\ Q_{out} = MeanField(Q_{in}) & \text{if } t = T) \end{cases} \tag{16}$$

where $Q = \{q_i^u\}$, $\{q_i^u\}$ denotes the probability of assigning label u to pixel i, Q_{in} denotes the input of one mean-field iteration, Q_{out} denotes the output Q of one mean-field iteration, Q_{final} denotes the final prediction results of CRF-RNN, P denotes the output of UNet, P_{norm} denotes the P that after softmax operation, t represents the t^{th} mean-field iteration, and T is the total number of mean-field iterations. Mean-field iteration [13] is considered as a stack of CNN layers which includes these steps: Message Passing, Re-Weighting, Compatibility transform, Adding Unary Potentials, and Normalization. In our study, the term $\Phi(y_i^u)$ in Equation (11) is the output of UNets and the term $\Psi(y_i^u, y_j^v)$is computed based on feature vectors of pixel i, j with information is derived from image features such as spatial location and

RGB values. The parameters of the Gaussian kernels $\theta_\alpha = 160, \theta_\beta = 3, \theta_\gamma = 3$ while w and μ are learned in the training phase, the RNN parameter iteration count T is set to $T = 10$ during the test time and $T = 5$ during the training time, according to [13].

By using the UNet that fully integrates CRF-RNN as layer on top, and making whole network possible to train end-to-end with the back-propagation algorithm, we can improve the polyp segmentation accuracy without offline post-processing for object delineation. The experiment's results in section V-A4 show the considerable increase in Dice score when using the CRF-RNN layer on top of all experimented networks.

IV. EXPERIMENTAL METHOD

A. DATASET

Several public available benchmark datasets are used for the training and evaluation of the proposed method. The examples are given from the datasets in Figure 5 Details of these datasets are summarized as below:

-*CVC-ClinicDB* dataset [15] consists of 612 images from 31 different types of polyps along with the corresponding ground truth masks of defined polyp regions. The ground truth masks are manually annotated by experts. All the images have a resolution of 384 × 288.

- *Kvasir-SEG* dataset [14], publicized by Simula Research Laboratory, includes 1000 polyp images with varying sizes from 332 × 482 to 1920 × 1072 and their corresponding ground truth masks manually annotated by expert endoscopists from Oslo University Hospital (Norway).

-*ETIS-Larib* dataset [16] contains 36 different types of polyps in 196 images with a resolution of 1225 × 966. These images were extracted from colonoscopy videos, and the ground truth masks were annotated by experts. This dataset is provided in the 2015 MICCAI automatic polyp detection sub-challenge as the test set.

-The *CVC-ColonDB* dataset [15] is contributed by the Machine Vision Group (MVG). This dataset consists of 300 polyp images and their corresponding pixel-level annotated polyp masks extracted from 15 video sequences. The images had a resolution of 574 × 500.

These datasets were obtained with different imaging systems. Each dataset contains binary masks as the ground truths to indicate the location of the polyps for each image. Expert endoscopists annotated all ground truths of polyp regions from the corresponding associated clinical institutions. There are similar image frames within a dataset. However, the datasets vary regarding the number of images, image resolution, availability, devices used for capturing, and the accuracy of the segmentation masks. In this work, we conduct experiments with different scenarios using training and testing data to compare the proposed model's performance over the SOTA approaches.

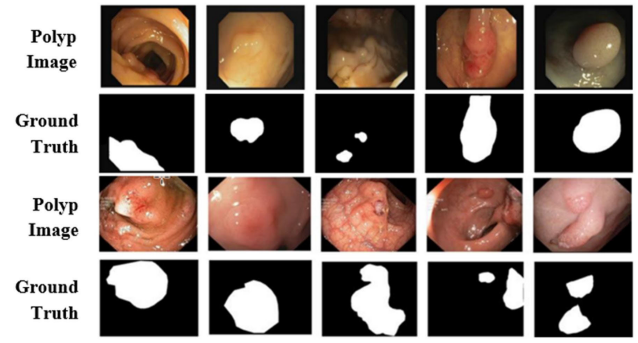

FIGURE 5. Examples of polyp segmentation datasets- The first and second lines are the polyp and mask images from the CVC-ClinicDB dataset; The third and fourth lines are the polyp and mask images from the Kvasir-SEG dataset.

B. DATA AUGMENTATION

One of the challenges in training polyp segmentation models is the insufficient numbers of data for training. Since the endoscopy procedures involving moving camera control, color calibrations are not consistent, the appearance of endoscopy images significantly changes across different laboratories. The data augmentation step extends endoscopy images into the space that can cover all their variances. By augmenting training data, we can also reduce the overfitting problem on training models. Figure 6 shows the examples of the data augmentation method applied to the original polyp image (a). The methods of augmentation used in our work include vertical flipping, horizontal flipping, random rotation between -10 and 10 degrees, random scaling ranging from 0.5 to 1.5, random shearing between -5 and 5 degrees, random Gaussian blurring with a sigma of 3.0, random contrast normalization by a factor of 1 to 1.5, random brightness ranging from 1 to 1.5, and random cropping and padding by 0–5% of height and width.

C. EVALUATION METRICS

For the evaluation of polyp segmentation, we use *Dice* coefficient as the main metric. Furthermore, to provide a general view of the effectiveness of our method, we also employed interception over union (*IoU*), recall (*Re*) which is also known as sensitivity, and precision (*Pre*). The evaluation metrics are calculated as follows:

$$Dice = \frac{2|PR \cap GT|}{|PR| + |GT|} \tag{17}$$

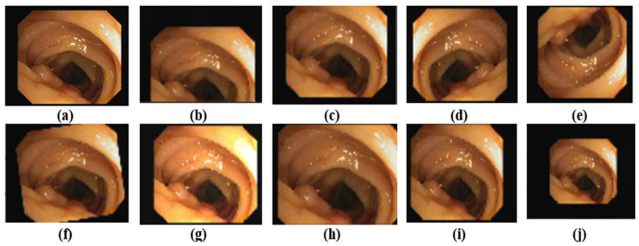

FIGURE 6. Examples of Data Augmentation: (a) the original polyp image, (b),(c) random shifting, (d) horizontal flipping, (e) vertical flipping, (f) random shearing, (g) random brightness, (h),(j) random scaling ranging, (i) random Gaussian blurring.

$$IoU = \frac{|PR \cap GT|}{|PR \cup GT|} \quad (18)$$

$$Re = \frac{|TP|}{|TP| + |FN|} \quad (19)$$

$$Pre = \frac{|TP|}{|TP| + |FP|} \quad (20)$$

where PR represents prediction results, GT is the ground-truth, TP is true positives, FP is false positives, and FN is false negatives. Metrics compute on every image, then average on the whole dataset across all images.

D. TRAINING SETUP

The proposed models are implemented using Keras and TensorFlow backend. All algorithms have been programmed/trained on a PC with a GeForce GTX 1080 Ti GPU. Weights pre-trained on ImageNet for encoders are used as initialization. The encoders are unfrozen, and the entire network is updated via Adam optimizer with the learning rate of $1e-4$ and the maximum epoch number of 500. The proposed loss function, combined asymmetric loss, is used for training models. The dataset is divided into batches with a mini-batch size of four for the training. The model generated at the epoch with max Dice score on the validation set is used as the final model.

V. EXPERIMENTS RESULTS AND ANALYSIS

A. ABLATION STUDY

To analyze the effect of each component in the proposed model on the segmentation performance, we performed an ablation study with model variants. To make equal to all ablation experiments, we conduct experiments on CVC-ClinicDB dataset. The dataset is split 80/10/10 for training, validation, and testing.

1) PERFORMANCE EVALUATION ON CNN PRE-TRAINED ENCODERS

We first evaluate UNets with different encoders. Several encoders are selected to evaluate their performance in polyp segmentation. The EfficientNet family from B0 to B7, MobileNetV2, ResNet variants, including ResNet18, ResNet34, ResNet101 have been used. Table 2 presents the overall results of the experiments. This table shows that EfficientNet family backbones significantly outperform ResNet and MobileNet in terms of Dice and IoU scores; EfficientNet backbones generally perform better as size increases; UNet-EfficientNetB7 gives the best segmentation performance with 93.72% Dice and 88.63% IoU.

2) THE EFFECT OF COMBINED ASYMMETRIC LOSS FUNCTION

Next, we evaluate the effect of the proposed loss function on models' performance and compare it with basic loss functions in polyp segmentation. We conducted experiments using three backbones, UNet-MobileNetV2, UNet-ResNet101, UNet- EfficientNetB7; the models are called UNet1, UNet2, and UNet3, respectively. We trained these models using binary cross-entropy loss (BCE loss), Dice loss, Asymmetric loss, and our proposed loss, i.e., combined asymmetric loss. The hyperparameters of loss functions are chosen for the best results of models with $\alpha = 0.4$ and $\beta = 1.6$. The improvements of performance metrics are reported in Table 3. This table demonstrates that our proposed loss function makes a better balance between precision and recall than other loss functions. Therefore, the performance of models trained with our proposed loss function is increased. Comparing to binary cross-entropy loss, the models trained by the proposed loss function could improve performance the most, specifically as follows: Unet1 (MobileNetV2 encoder) could improve Dice by 6.23% and IoU by 4.75%; Unet2 (ResNet101 encoder) could improve Dice by 3.37% and IoU by 1.6%; Unet3 (EfficientNetB7 encoder) could improve Dice by 3.37% and IoU by 1.6%. Although the precision may be decreased, the proposed loss function can make a trade-off between precision and recall so that the Dice score can be increased. Figure 7 illustrates the Dice scores of models trained by cross-entropy loss, Dice loss, asymmetric loss, and the proposed loss. This figure shows that the Dice scores of models trained by the proposed loss function outperform the others. Moreover, Figure 8 describes the effects on the network learning progress of the proposed loss function (combined asymmetric loss) and the cross-entropy loss function. This figure shows that the validation loss values are less variable during training when the model is trained by our proposed loss function than when the model is trained by the cross-entropy loss function.

TABLE 2. Comparison of UNet models with different backbones.

Network	Dice (%)	IoU (%)	Re (%)	Pre (%)
UNet-MobileNetV2	88.72	81.56	90.85	89.36
UNet-ResNet18	74.05	63.2	74.14	79.05
UNet-ResNet34	79.81	69.12	78.67	84.81
UNet-ResNet50	77.43	64.98	80.01	77.55
UNet-ResNet101	79.15	70.63	82.19	86.43
UNet-EfficientNetB0	86.14	76.66	82.84	91.18
UNet-EfficientNetB1	92.27	86.94	92.24	94.1
UNet-EfficientNetB2	91.77	85.4	93.97	90.42
UNet-EfficientNetB3	90.02	82.54	91.63	89.93
UNet-EfficientNetB4	92.06	86.57	91.65	94.28
UNet-EfficientNetB5	92.42	86.5	93.36	92.44
UNet-EfficientNetB6	92.87	87.94	**95.05**	92.45
UNet-EfficientNetB7	**93.72**	**88.63**	93.34	**94.56**

3) THE EFFECT OF TRANSFER LEARNING

This work adopts a transfer learning approach with UNet architecture for polyp segmentation by using CNN models pre-trained on the ImageNet dataset as the encoder. To evaluate the effect of this transfer learning method, we train UNet from scratch and compare the received results with the result from the transfer learned UNet. We conducted experiments using six backbones: Unet-MobileNetV2, Unet-Resnet50, Unet-Resnet101, Unet-EfficientNetB5, Unet-EfficientNetB6, Unet-EfficientNetB7. The comparations of performance metrics for polyp

TABLE 3. Peformance of Unets trained with different loss functions.

Network	Dice (%)	IoU (%)	Re (%)	Pre (%)
UNet1 with BCE loss	82.49	76.81	83.15	90.63
UNet1 with Dice loss	85.61	75.95	79.77	**93.87**
UNet1 with asymmetric loss	87.78	79.48	85.84	92.22
UNet1 with proposed loss	**88.72**	**81.56**	**90.85**	**89.36**
UNet2 with BCE loss	57.89	47.97	54.49	82.52
UNet2 with Dice loss	59.82	43.58	47.29	85.30
UNet2 with asymmetric loss	72.51	58.7	73.03	76.89
UNet2 with proposed loss	**79.15**	**70.63**	**82.19**	**86.43**
UNet3 with BCE loss	90.35	87.03	91.64	**94.44**
UNet3 with Dice loss	90.05	82.20	91.20	89.79
UNet3 with asymmetric loss	92.72	86.64	91.79	92.72
UNet3 with proposed loss	**93.72**	**88.63**	**94.56**	93.34

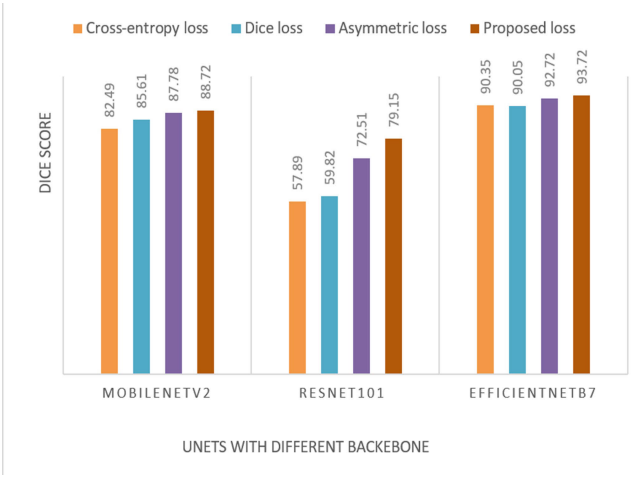

FIGURE 7. Dice scores of models trained by difference loss functions.

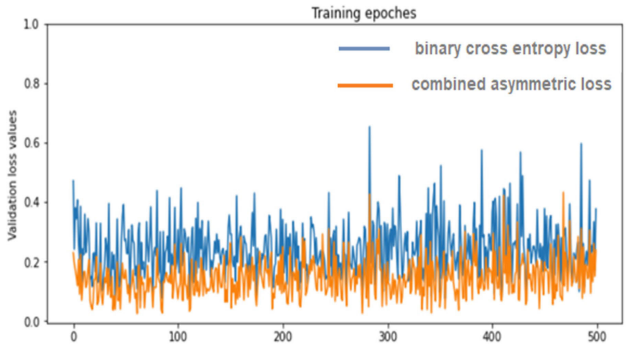

FIGURE 8. The effect of the proposed loss function on network learning progress on the same dataset compared to the cross-entropy loss function.

segmentation between the UNet trained from scratch and transfer learning methods are reported in Table 4. The table demonstrates that the performance of models trained by the transfer learning method is significantly improved compared to those trained from scratched. In addition, when the models are deeper, the performance improvement is greater.

4) THE EFFECT OF CONDITIONAL RANDOM FIELDS AS RECURRENT NEURAL NETWORK LAYER

We adapted some experiments to test whether using a CRF-RNN layer on top of the polyp segmentation networks

TABLE 4. Comparison of Unet models trained from scratch and transfer learning.

Model	Transfer learning		Trained from scatch	
	Dice	IoU	Dice	IoU
UNet-MobileNetV2	88.72	81.56	87.51	79.41
UNet-Resnet50	77.43	64.98	73.61	66.07
UNet-Resnet101	79.15	70.63	72.08	62.43
UNet-EfficientNetB5	92.42	86.5	89.42	82.31
UNet-EfficientNetB6	92.87	87.94	88.12	81.08
UNet-EfficientNetB7	93.72	88.63	89.75	82.81

improved the segmentation quality. These experiments aim to compare the performance difference between using and not using a CRF-RNN layer on top of the segmentation network. Underlying network architectures used for polyp segmentation are several UNets with different backbones, including UNet-MobileNetV2, UNet-ResNet101, and UNet-EfficientNetB7. The results are presented in Table 5. This table shows a considerable increase in Dice score when using a CRF-RNN layer on top of all experimented networks. More specifically, the UNet-EfficientNetB7 with a CRF-RNN layer on top achieves the most improvements of 1.83% in terms of Dice score, and UNet-MobileNetV2 with CRF-RNN increases the least by 0.92% in terms of Dice, as can be calculated from average metrics in Table 5. The improvement in results demonstrates the advantage of using a CRF-RNN layer on top of segmentation networks. Moreover, Figure 9 illustrates the comparison of Dice scores of UNets with and without a CRF-RNN layer on top. This figure also shows improvements in Dice score when using a CRF-RNN layer on top of all experimented networks.

Finally, some examples of different segmentations produced by model variants are depicted in Figure 10. The figure describes the UNet model with EfficientNetB7 backbone and CRF-RNN trained by combined asymmetric loss function can recognize the polyp mask more accurately than other models. This figure also shows that our model has the robustness with the ability to detect polyps on challenging images (e.g., blurriness, low-quality images in Figure 10(a), 10(b), small polyps in 10(e)) and perform well on easier images (e.g., Figure 10(c), 10(d)).

B. COMPARISON TO EXISTING METHODS

This section compares our proposed CRF-EfficientUNet to several recent SOTAs for polyp segmentation. Results for the compared models are reported in their respective papers. From the previous ablation study, we select the UNet-EfficientNetB7 with combined asymmetric loss function and CRF-RNN layer as the comparison model for this section. Then, we conduct experiments with different scenarios of training and testing data. The hyperparameters of asymmetric loss function are chosen based on the empirical evaluation, with $\alpha = 0.4, \beta = 1.6$ on the CVC-Clinic dataset, and $\alpha = 0.3, \beta = 1.3$ on the Kvasir-SEG dataset. We present and compare the results of the proposed method with existing methods in terms of learning ability, generalization capability on the same dataset, and cross-dataset.

TABLE 5. The effect of CRF-RNN layer.

Network	Dice (%)	IoU (%)	Re (%)	Pre (%)
UNet-MobileNetV2 without CRF-RNN	88.72	81.56	**90.85**	89.36
UNet-MobileNetV2 with CRF-RNN	**89.64**	**81.65**	87.53	**92.46**
UNet-ResNet101 without CRF-RNN	79.15	**70.63**	**82.19**	86.43
UNet-ResNet101 with CRF-RNN	**80.55**	68.99	76.87	**87.04**
UNet-EfficientNetB7 without CRF-RNN	93.72	88.63	93.34	94.56
UNe-EfficientNetB7 with CRF-RNN	**95.55**	**92.23**	**96.31**	**95.69**

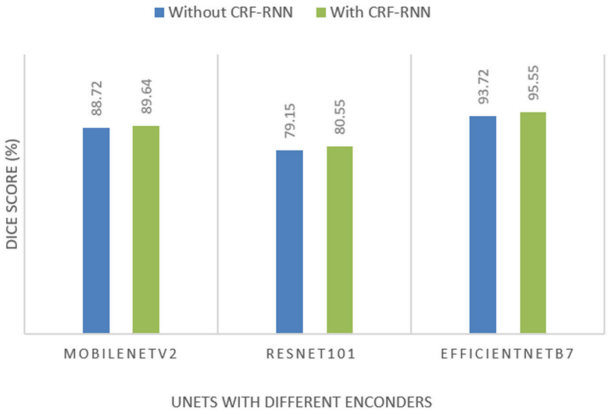

FIGURE 9. Comparison Dice scores of models with and without CRF-RNN.

1) RESULTS ON THE SAME DATASETS

We conduct two experiments to validate the model's learning ability when the training and test set are from the same dataset. The first experiment uses CVC-Clinic dataset, and the second uses Kvasir-SEG dataset. These experiments are conducted with a five-fold cross-validation scheme. In this scheme, four folds are used for training, while the remaining fold is used to evaluate performance. The training and evaluating processes are repeated five times, and the mean values of the evaluation metrics are reported. The results are compared with several SOTAs. Table 6 and Table 7 show the comparisons of the quantitative results on CVC-Clinic and Kvasir-SEG, respectively. As shown in these tables, our method outperforms all other methods in Dice and IoU metrics across both datasets. Specifically, Table 5 shows that our proposed methods achieve the best performance on CVC-Clinic dataset with Dice of 95.12% and IoU of 91.85%, outperforming the second-best ResUNet++ CRF by 3.09% in Dice and 2.87% in IoU. In Table 6, on Kvasir-SEG dataset, our proposed method also gets the highest Dice of 92.72% and the second-highest IoU of 87.69% (the highest is Efficient UNet multi-scale attention with IoU of 88.69%). These results demonstrate that our model has a strong learning ability to segment polyps effectively.

2) RESULTS ON CROSS-DATASET

We carry out experiments with training and testing across different datasets to measure the generalization capability of the proposed method. Since different polyp datasets have different image properties and feature distributions, the models need to generalize well to have good performance. In this session, we train models on CVC-ClinicDB, Kvasir-SEG, and a mixed Kvasir and CVC-ClinicDB, respectively, and use the other independent datasets: ETIS-Larib, CVC-ColonDB for testing. Then, we compare the results with current works that have the same training and testing data scenarios. The results are reported in comparison tables, where 'n/a' denotes unavailable results, and '*' indicates the results generated using the released code.

TABLE 6. Comparison of quantitative results on CVC-ClinicDB dataset.

Network	Dice (%)	IoU (%)	Re (%)	Pre (%)
UNet [8]	71.47	43.34	63.06	92.22
DoubleU-Net [26]	92.39	86.11	84.57	95.92
UNet++ [27]	78.15	72.41	80.64	90.76
ResUNet++ [28]	91.99	88.92	93.91	84.45
A-DenseUNet [35]	89.12	85.53	94.48	92.66
PolypSegNet [32]	84.79	78.32	84.34	95.75
ResUNet++CRF [34]	92.03	88.98	93.93	84.59
PraNet [25]	89.9	84.9	n/a	n/a
Proposed	**95.12** ±0.92	**91.85** ±1.57	**97.94** ±0.75	**96.83** ±1.35

TABLE 7. Comparison of quantitative results on Kvasir-SEG dataset.

Network	Dice (%)	IoU (%)	Re (%)	Pre (%)
UNet [8]	71.47	43.34	63.06	92.22
UNet++ [27]	80.21	72.15	79.14	93.21
ResUNet++ [28]	81.19	80.68	85.78	77.42
A-DenseUNet [35]	90.85	86.15	94.48	97.66
PolypSegNet [32]	88.72	82.86	91.68	92.54
SegNetImproved [31]	n/a	87.2	96.6	85.0
PraNet [25]	89.8	84.0	n/a	n/a
Efficient U-Net [42]	87.85	88.69	n/a	n/a
multi-scale attention [42]	±0.11	±0.63	n/a	n/a
Proposed	**92.72** ±2.23	**87.69** ±1.73	**97.02** ±2.23	**94.92** ±2.6

First, we train the model with CVC-ClinicDB dataset. Table 8 presents the results and comparison with several SOTAs for polyp segmentation. On ETIS-Larib test set, the proposed method gives the best segmentation performance with 79.37% Dice, 68.65% IoU, recall of 79.44%, and precision of 80.07%. The proposed method obtains the best results on CVC-ColonDB test set: 86.8% Dice, 77.43% IoU, recall of 86,4%, and precision of 85.52%. These results indicate that our method outperforms other SOTAs on both test sets. Especially with the CVC-ColonDB test set, our Dice score is 12.1% higher than PolypSegNet's [31], which is the second-highest method. In addition, Figure 11 presents examples of segmentations produced by the proposed model with challenging images of the ETIS-Larib dataset. This figure also demonstrates that our model has the robustness to detect polyps on challenging images such as varying shapes and textures of polyps, small polyps, and the presence of image artifacts like saturation, artifact, bubbles, intestinal contents.

Next, we train the model with Kvasir-SEG dataset. Table 9 shows the results and comparison with other models for

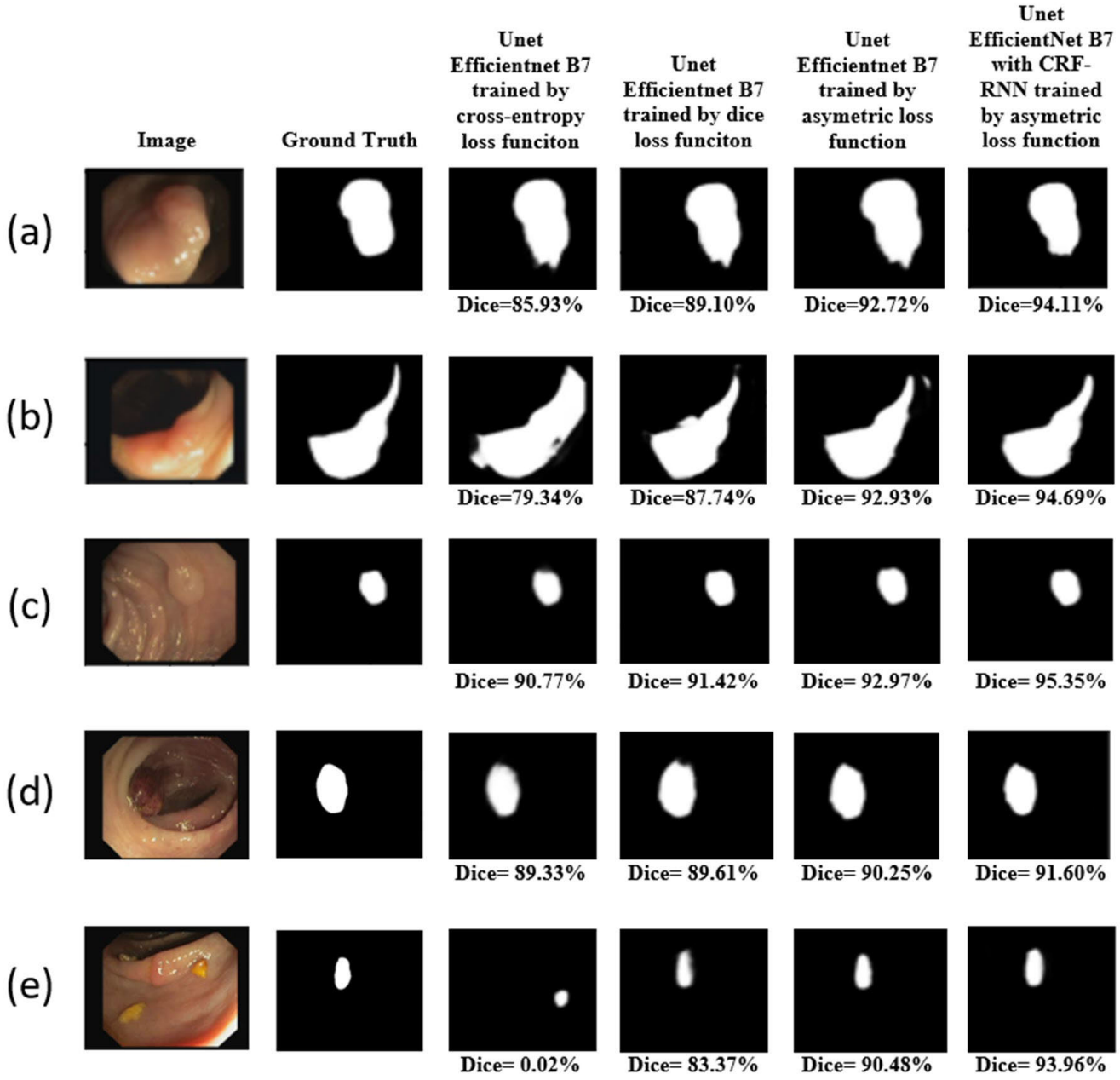

FIGURE 10. Examples of different segmentations produced by model variants with the ability of the system to perform well on both easy and challenging images: (a),(b) low-quality images, (c),(d) easier images, (d) horizontal flipping, (e) small polyps.

polyp segmentation. Like the previous experiment with CVC-ClinicDB, our proposed method also outperforms all other methods on both test sets. On ETIS-Larib test set, we obtain the best segmentation performance with 78.53% Dice, 66.95% IoU. On CVC-ColonDB test set, the proposed method gets the best results with 85.59% Dice, IoU 76.19%, recall of 88,07%, and precision of 86.78%, outperforms the second-highest method ResUNet++ TTA [33] by 29.63%

Finally, we use 1450 images, including 900 images in Kvsir-SEG and 550 images in CVC-ClinicDB for training models. Table 9 presents the results of the cross-data generalizability of methods. The table shows that our proposed method achieves the highest results on both test sets with 78.35% Dice on ETIS-Larib and 86.04% Dice on CVC-ColonDB. We have compared the results with the existing works that used the same scenarios of using training and test data. Our method also outperforms all in both Dice and IoU metrics. Especially with the CVC-ColonDB test set, our Dice score is 15.14% higher than the second-highest method PraNet [24].

In this section, we conduct experiments to measure the generalization capability of the proposed method.

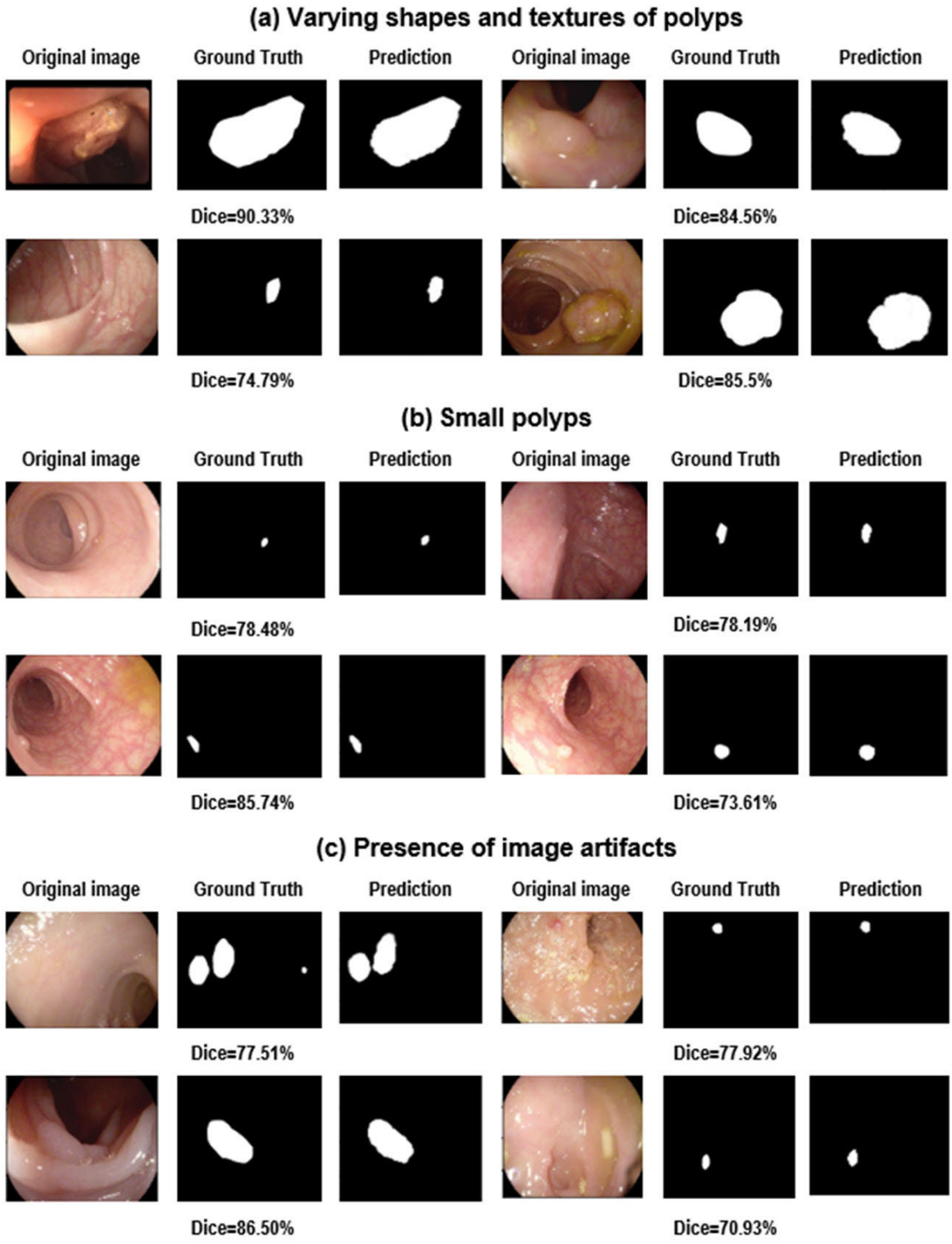

FIGURE 11. Examples of segmentations produced by the proposed model with challenging images of the ETIS-Larib dataset.

Generalization capability checks the usefulness of the model across different available datasets coming from different hospitals. A good generalizable model could be a significant step toward developing a good clinical system. It should be noted that the performance of the proposed method outperforms all SOTAs across independent test sets in terms of Dice metric.

TABLE 8. Comparison results on cross-dataset using Clinic-DB as the training set.

	Network	Dice (%)	IoU (%)	Re (%)	Pre (%)
ETIS-Larib	UNet [8] *	57.25	n/a	n/a	n/a
	UNet++ [27] *	55.12	n/a	n/a	n/a
	ResUNet++ [28]	40.12	63.98	42.32	40.13
	ResUNet++ TTA [34]	40.27	65.22	39.69	42.25
	ResNet101-Mask-RCNN [21]	70.42	61.34	72.59	80.0
	Ensemble Mask-RCNNs [22]	n/a	66.07	74.37	66.07
	DoubleU-Net [26]	76.49	62.55	71.56	80.07
	PolypSegNet [32]	68.6	n/a	n/a	n/a
	Proposed	**79.37**	**68.85**	**79.44**	**80.07**
CVC-ColonDB	UNet [8]*	65.32	n/a	n/a	n/a
	UNet++ [27]*	61.85	n/a	n/a	n/a
	ResUNet++ [28]	54.89	69.42	55.77	58.16
	ResUNet++ TTA [34]	56.86	70.8	57.02	59.35
	Ensemble Mask-RCNNs [22]	n/a	69.46	77.92	76.25
	DoubleU-Net [26] *	71.21	n/a	n/a	n/a
	PolypSegNet [32]	74.7	n/a	n/a	n/a
	Proposed	**86.8**	**77.43**	**86.4**	**85.52**

TABLE 9. Comparison results on cross-dataset using Kvasir-SEG as the training set.

Method	ETIS-Larib		CVC-ColonDB	
	Dice	IoU	Dice	IoU
UNet [8] *	60.25	n/a	66.12	n/a
UNet++ [27] *	58.43	n/a	65.21	n/a
ResUNet++ [28]	40.17	64.15	51.35	67.42
ResUNet++ TTA [34]	40.14	64.68	55.93	70.3
DoubleU-Net [26] *	64.4	n/a	75.3	n/a
PolypSegNet [32]	71.8	n/a	n/a	n/a
Proposed	**78.53**	**66.95**	**85.56**	**76.19**

TABLE 10. Comparison results on cross-dataset using mixed Kvasir-SEG and CVC-ClinicDB dataset as the training set.

Method	ETIS-Larib		CVC-ColonDB	
	Dice	IoU	Dice	IoU
UNet [8]*	39.0	33.5	52.12	44.4
UNet++ [27]*	40.1	34.4	43.83	41.0
ResUNet++ [28]	n/a	n/a	49.74	68.0
ResUNet++ TTA [34]	n/a	n/a	50.84	68.59
PraNet [25]	62.8	56.7	70.9	64.0
DoubleU-Net [26]*	45.4	n/a	52.5	n/a
Proposed	**78.35**	**67.03**	**86.04**	**77.22**

These results indicate that the proposed method has better generalizability than existing methods, and it can be a compelling choice for practical applications with considerable data variations.

VI. CONCLUSION

This paper proposes CRF-EfficientUNet, an improved UNet framework for polyp segmentation. We present a novel UNet-based architecture extended from UNet with the EfficientNet B7 encoder and the CRF-RNN layer on top. A novel loss function is proposed for training CRF-EfficientUNet to solve the unbalanced data problem and achieve better performance. Besides, we use the transfer learning method to train and validate the proposed method on various datasets, i.e., Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, EITS-Larib, with different scenarios of using training and test data. Moreover, we check the generalization capability of the proposed method by training the proposed model on Kvasir-SEG and CVC-ClinicDB and testing it over other independent datasets: ETIS-Larib, CVC-ColonDB. The results of the proposed method outperform all SOTAs on the same dataset and cross-dataset. These results indicate that our proposed method has better generalizability and learning ability than others. In the future, we will focus on reducing the network size with better performance to build a model which can be an effective choice for practical automated polyp segmentation. Besides, the proposed method can be converted to 3D models and easily applied to other screening modalities like CT and MRI.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] J.-F. Rey and R. Lambert, "ESGE recommendations for quality control in gastrointestinal endoscopy: Guidelines for image documentation in upper and lower GI endoscopy," *Endoscopy*, vol. 33, no. 10, pp. 901–903, Sep. 2001.

[3] M. Gschwantler and S. E. A. Kriwanek, "High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics," *Eur. J. Gastroenterol. Hepatol.*, vol. 14, no. 2, pp. 183–188, 2002.

[4] A. Leufkens, M. Van Oijen, F. Vleggaar, and P. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[5] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, pp. 1–9, Jul. 2017.

[6] S. Ali, F. Zhou, C. Daul, B. Braden, A. Bailey, S. Realdon, J. East, G. Wagnières, V. Loschenov, E. Grisan, W. Blondel, and J. Rittscher, "Endoscopy artifact detection (EAD 2019) challenge dataset," 2019, *arXiv:1905.03209*.

[7] T. Ross *et al.*, "Robust medical instrument segmentation challenge 2019," 2020, *arXiv:2003.10299*.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[11] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[12] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. New York, NY, USA: Curran Associates, 2011, pp. 109–117.

[13] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. Cham, Switzerland: Springer, Dec. 2015, pp. 1529–1537.

[14] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modeling*. Springer, 2020, pp. 451–462.

[15] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014.

[16] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[17] L. T. Thu Hong, N. Chi Thanh, and T. Q. Long, "Polyp segmentation in colonoscopy images using ensembles of U-Nets with EfficientNet and asymmetric similarity loss function," in *Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Oct. 2020, pp. 1–6.

[18] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 8, pp. 2144–2151, Aug. 2012.

[19] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "A classification-enhanced vote accumulation scheme for detecting colonic polyps," in *Proc. Int. MICCAI Workshop Comput. Clin. Challenges Abdominal Imag.*, Berlin, Germany: Springer, 2013, pp. 53–62.

[20] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor CNN always perform better?" in *Proc. 13th Int. Symp. Med. Inf. Commun. Technol. (ISMICT)*, May 2019, pp. 1–6.

[21] J. Kang and J. Gwak, "Ensemble of instance segmentation models for polyp segmentation in colonoscopy images," *IEEE Access*, vol. 7, pp. 26440–26447, 2019.

[22] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov, "Fully convolutional neural networks for polyp segmentation in colonoscopy," *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101340F.

[23] L. Zhang, S. Dolwani, and X. Ye, "Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons," in *Proc. Annu. Conf. Med. Image Understand. Anal.* Cham, Switzerland: Springer, 2017, pp. 707–717.

[24] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273.

[25] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[27] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 225–2255.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[29] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang, and X. Yang, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 741–748, 2018.

[30] H. M. Afify, K. K. Mohammed, and A. E. Hassanien, "An improved framework for polyp image segmentation based on SegNet architecture," *Int. J. Imag. Syst. Technol.*, vol. 31, no. 3, pp. 1741–1751, Sep. 2021.

[31] T. Mahmud, B. Paul, and S. A. Fattah, "PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104119.

[32] Y. Fang, C. Chen, Y. Yuan, and K.-Y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 302–310.

[33] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. D. Johansen, P. Halvorsen, and M. A. Riegler, "A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2029–2040, Jun. 2021.

[34] S. Safarov and T. K. Whangbo, "A-DenseUNet: Adaptive densely connected UNet for polyp segmentation in colonoscopy images with atrous convolution," *Sensors*, vol. 21, no. 4, p. 1441, Feb. 2021.

[35] C. Yang, X. Guo, M. Zhu, B. Ibragimov, and Y. Yuan, "Mutual-prototype adaptation for cross-domain polyp segmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3886–3897, Oct. 2021.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[37] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2019.

[38] N. Nguyen and S.-W. Lee, "Robust boundary segmentation in medical images using a consecutive deep encoder-decoder network," *IEEE Access*, vol. 7, pp. 33795–33808, 2019.

[39] L. F. Sánchez-Peralta, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, "Unravelling the effect of data augmentation transformations in polyp segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 12, pp. 1975–1988, Dec. 2020.

[40] N.-Q. Nguyen, D. M. Vo, and S.-W. Lee, "Contour-aware polyp segmentation in colonoscopy images using detailed upsampling encoder-decoder networks," *IEEE Access*, vol. 8, pp. 99495–99508, 2020.

[41] S. Poudel and S.-W. Lee, "Deep multi-scale attentional features for medical image segmentation," *Appl. Soft Comput.*, vol. 109, Sep. 2021, Art. no. 107445.

LE THI THU HONG received the bachelor's degree from the Hanoi University of Science and Technology, Vietnam, in 2003, and the master's degree from Le Quy Don University, Vietnam, in 2010. Currently, she is a Researcher with the Institute of Information Technology, AMST, Hanoi, Vietnam. Her research interests include image processing, computer vision, medical image analysis, and deep learning.

NGUYEN CHI THANH received the Ph.D. degree in computer science from the Nagaoka University of Technology, Japan, in 2012. He is currently a Researcher with the Institute of Information Technology, AMST, Hanoi, Vietnam. His research interests include deep learning, computer vision, medical image analysis, and natural language processing.

TRAN QUOC LONG received the Ph.D. degree in computer science from the Georgia Institute of Technology, USA, in 2015. He is currently a Lecturer with the Faculty of Information Technology, University of Technology, Hanoi National University, Vietnam. His research interests include deep learning, computer vision, medical image analysis, and natural language processing.